Segregation of whispered speech interleaved with noise or speech maskers

Nandini Iyer[1], *Douglas S. Brungart*[2], *& Brian D. Simpson*[1]

[1] Air Force Research Laboratory, Wright-Patterson Air Force Base, OH
[2] Walter Reed Army Medical Center, Bethesda, MD

Nandini.Iyer@wpafb.af.mil, Douglas.Brungart@us.army.mil, Brian.Simpson@wpafb.af.mil

Abstract

Some listening environments require listeners to segregate a whispered target talker from a background of other talkers. In this experiment, a whispered speech signal was presented continuously in the presence of a continuous masker (noise, voiced speech or whispered speech) or alternated with the masker at an 8-Hz rate. Performance was near ceiling in the alternated whisper and noise condition, suggesting that harmonic structure due to voicing is not necessary to segregate a speech signal from an interleaved random-noise masker. Indeed, when whispered speech was interleaved with voiced speech, performance decreased relative to the continuous condition when the target talker was voiced but not when it was whispered, suggesting that listeners are better at selectively attending to unvoiced intervals and ignoring voiced intervals than the converse.

Index Terms: Target intelligibility, whispered speech, voiced speech, sequential segregation, simultaneous segregation.

1. Introduction

Whispering is an effective and efficient form of communication that is often adopted in covert operations in order to restrict the range over which the talker can be heard. Whispered speech is produced by modulating the flow of air through partially open vocal folds. Because the source of excitation is turbulent air flow, the acoustic characteristics of whispered speech differs from voiced speech [1, 2]. Despite the acoustic differences, whispered speech conveys much of the same information as voiced speech. Studies have shown that talkers can convey not just phonetic information [3, 4], but also information about talker gender [5] as well as listener emotional state [6]. And while there is an abundance of studies on the production and perception of whispered speech, very few studies have investigated whether whispered speech can be easily segregated from a background of interfering talkers. Segregating target talker whispered speech is an important skill for talkers who wish to communicate confidentially in noise, and it may also have implications for listeners who are attempting to parse complex auditory scenes with the modulated noise-like signals provided by cochlear implants.

Two studies investigated the segregation of simultaneously presented whispered vowels [7, 8] in a standard double vowel identification paradigm. Both experiments found that pairs of concurrent whispered vowels were identified at the same rate as two vowels with the same fundamental frequency. When a difference was introduced between the concurrent vowel pairs (i.e., voiced vs. whisper), there was about a 10% improvement in identification of the vowel pairs over the condition where both were whispered. Also, in a combination, the whispered vowel in a voiced/whispered pair was identified significantly better than in listening conditions with two whispered vowels. In a more recent study, [9] investigated the importance of voicing in the recognition of concurrent speech signals. They showed that, when audibility was accounted for, listeners' performance were comparable when they indentified a voiced target from a voiced masker, a voiced target from a whispered masker, or a whispered target from a voiced masker. However, performance declined in conditions where both target and masking talkers were whispered.

A recent study [10] also explored whether the lack of temporal fine structure in whispered speech led to any differences in a listener's ability to make use of temporal fluctuations in a noise masker compared to a steady-state masker. They observed that listeners derived considerable benefits from fluctuating maskers and even though the overall recognition of whispered speech is lower, the amount of masking release obtained with whispered speech was comparable to normal speech.

While in most communication situations, listeners have to integrate information about the target from spectro-temporal glimpses of masking and target talkers presented simultaneously, some scenarios require integrating target signals across time because it is interrupted by brief periods of silence or noise (such as cell-phone or radio communication). And while it is clear that listeners can segregate two talkers based on perceptual differences (such as pitch differences, voicing differences, etc.) between them in a simultaneous task, it is not clear if they can also segregate two talkers presented sequentially. To our knowledge, there is no data available on listeners' ability to segregate sequentially presented whispered speech compared to voiced speech. In a related study, [11] showed that when an interrupted speech target was alternated with a noise masker, performance generally improved over listening conditions when both were presented simultaneously. However, when two interrupted speech signals were alternated, performance declined relative to the continuous presentation condition. One possible explanation for this result is that listeners in the alternated speech and noise condition use the contrast between the periodic temporal structure of voiced speech and the random temporal structure of noise to segregate the time intervals associated with the speech target and noise masker. It is not clear if there would be an improvement in intelligibility when a speech masker is rendered more noise-like, such as with whispered speech. The goal of the current experiment was to measure target intelligibility (either voiced or whispered target phrases) in the presence of a whispered, voiced or noise masker when the two were alternated with each other or presented simultaneously.

2. Methods

2.1. Listeners

16 listeners (9 males, 7 females) participated in the study. The listeners ranged in age from 21-25 years. All listeners had normal audiometric thresholds (<20 dB at octave frequencies between 250-8000 Hz). All subjects were well-practiced in speech perception tasks, signed an informed consent

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Segregation of whispered speech interleaved with noise or speech maskers	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory,Wright-Patterson AFB,OH,45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Some listening environments require listeners to segregate a whispered target talker from a background of other talkers. In this experiment, a whispered speech signal was presented continuously in the presence of a continuous masker (noise voiced speech or whispered speech) or alternated with the masker at an 8-Hz rate. Performance was near ceiling in the alternated whisper and noise condition, suggesting that harmonic structure due to voicing is not necessary to segregate a speech signal from an interleaved random-noise masker. Indeed, when whispered speech was interleaved with voiced speech, performance decreased relative to the continuous condition when the target talker was voiced but not when it was whispered, suggesting that listeners are better at selectively attending to unvoiced intervals and ignoring voiced intervals than the converse.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

document, and were compensated for their participation in the study.

2.2. Speech Material

Target and masker phrases were recordings of sentences from the Coordinate Response Measure (CRM) corpus [12]. The phrases in the corpus are of the form "Ready [call sign], go to [color], [number] now. Eight possible call signs (Arrow, Baron, Charlie, Eagle, Hopper, Laker, Ringo, Tiger), four possible colors (white, blue, green and red) and eight possible numbers (1-8) result in 256 unique phrases per talker. Voiced and whispered versions of the corpus were recorded from eight talkers (four males, four females). The recordings were made at a 32 kHz sampling rate with a B&K 2131 microphone mounted on a stand positioned directly in front of the talker in a sound-treated double-walled audiometric chamber.

Talker Configuration	Designation
Fixed	
Target Voiced – Masker Voiced	TS
Target Whisper – Masker Whisper	T(W)S(W)
Target Voiced – Noise	TN
Target Whisper – Noise	T(W)N
Mixed	
Target Whisper – Masker Voiced	T(W)S
Target Voiced – Masker Whisper	TS(W)

Table 1: *Talker configuration and corresponding designations in the experiment for alternating and simultaneous conditions*

Target sentences, denoted by the call sign 'Baron', were presented to listeners in the presence of two kinds of maskers: noise or speech. When the masker was noise, the target sentence was multiplied in the frequency domain with a Gaussian noise and inverse Fourier transformed prior to presentation. This resulted in a noise stimuli that had a spectral shape identical to the target speech, but completely unintelligible. The noise masker was scaled so that it was 8 dB more intense than the target speech signal. When the masker was speech, a CRM phrase with a different call sign, color and number than the target phrase was selected. In order to maximize target-masker similarity, target and masker phrases were always selected so that the sex of the two talkers was the same within a trial. Listeners heard voiced and whispered versions of the target and/or masking talkers (for speech maskers). The specific talker configurations and designations are shown on Table 1. In addition, the target and masking phrases were either interrupted or continuous. When the target phrases were interrupted, both the target and masker were multiplied by a square wave with an 8 Hz interruption rate. The target signal was multiplied by the 'on' phase and the masker was multiplied by the 'off' phase of the square wave. The two resulting waveforms were then combined, resulting in a stimulus that alternated between the target and masker phrases. The stimuli used in the alternating condition are depicted in Figure 1. When the target was continuous, its onset was simultaneous with an uninterrupted masker. All target and masker phrases were scaled so that the target-to-masker ratio (TMR) varied from 8 to -20 dB in 4 dB steps.

2.3. Procedure

The speech stimuli were presented diotically via Beyerdyanamic DT990 Pro headphone. Listeners were seated in front of a computer monitor in a sound-treated room and responded to the target signal using a mouse. Response choices were displayed to listeners in a 4×8 matrix of colored-digits. Listeners chose the target color-number combination by moving their cursor and selecting a colored-number on the display. Correct response feedback was provided on every trial. Within a block of trials, the talker configuration, signal-to-noise ratio, and presentation mode (continuous vs. whispered) was randomly selected for every listener.

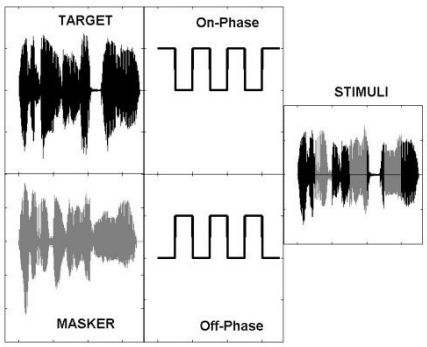

Figure 2: *Depiction of target (black) & masker waveform (gray), multiplied by the on & off phase of a square-wave respectively, resulting in an alternating signal comprising of target & masker segments.*

3. Results & Discussion

3.1. Validation of voiced and whispered corpus

An initial validation of voiced and whispered corpus was made in order to verify if trends obtained using the publicly available CRM corpus was similar to those obtained with the recorded corpus in the current study. Figure 2 depicts proportion correct color and number responses obtained by listeners when the masking talker was either the same talker (TT), different same sex talker (TS) or a different sex talker (TD) compared to the target. The left panel shows performance in listening conditions where the target and masking talkers were voiced, whereas the right panel depicts performance with two whispered talkers.

Overall, listener performance with the voiced corpus

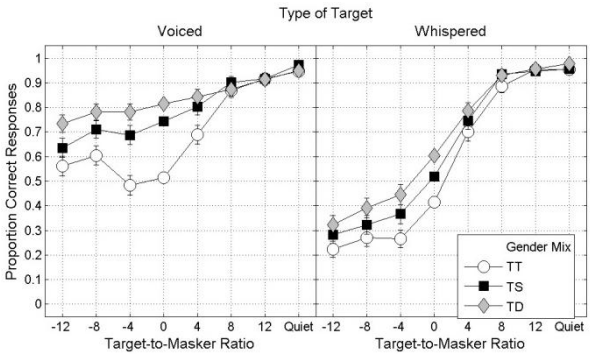

Figure 2: *Proportion correct target identification as a function of TMR for voiced (left panel) or whispered (right panel) corpus. The circles represent performance for listeners in conditions where the masker is the same talkers (TT) as the target, the squares represents performance in a condition when the masker is the same-sex talker (TS) as the target, and diamonds show performance in conditions when the target and masker are two different sex talkers (TD). The error bars represent the 95% confidence intervals around the mean.*

is similar to that obtained with the publicly available CRM corpus. Specifically, listeners are able to use level cues when segregating a target speech signal from a same talker masker [13]. Performance was best when the target speaker was different from the masker and decreased as the similarities increased. Further, performance declined only slightly over a 12 dB range (from 0 to -12 dB). For whispered speech, while the listeners could still segregate a target from a perceptually similar masker, they could no longer listen for the quieter target. Also, performance in all three configurations decreased more rapidly with decreasing TMR with whispered speech than with voiced speech. The validation study suggests that, while whispered speech lacks the pitch cue that tends to dominate gender perception in voiced speech, different from voiced speech, it contains enough information to segregate talkers based on sex differences. Indeed, [5] showed that listeners can accurately identify the sex of a talker when isolated vowels are presented.

3.2. Effect of talker configuration

Figure 3 depicts proportion correct color-number responses as a function of TMR in conditions when a noise masker was presented simultaneously (left panel) or alternated (right panel) with a voiced (circles) or whispered (squares) target. The TMR ranges also reflect the fact that the noise masker was scaled by 8 dB compared to the target signal. As previously reported by [10], performance is near-ceiling in conditions when a target speech signal, voiced or whispered, is alternated with a noise masker. From the figure, it is clear that there are no differences in performance between the two types of target talkers in noise.

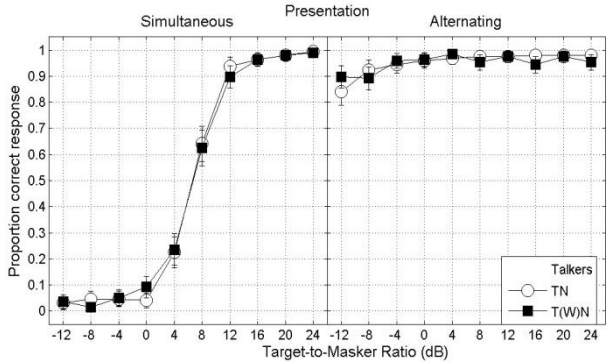

Figure 3: *Proportion correct color number responses for a target phrase, voiced (circles) or whispered (squares) presented simultaneously (left panel) or alternating (right panel) with a noise masker.*

Figure 4 depicts performance in listening configurations where target intelligibility was measured in the presence of a speech interferer as a function of TMR. The left panel depicts performance in listening conditions where the target and masker were presented simultaneously, whereas the right panel depicts performance in the alternating condition. The circles and squares depict performance in fixed listening conditions when both talkers were voiced or whispered respectively. The diamonds depict performance in mixed listening conditions.

From the figure, it is apparent that target identification was consistently most difficult when the target and masker were both whispered (black squares). When the target and masking phrases were voiced, performance improved compared to the whispered condition. The largest improvement with was obtained in the simultaneous listening condition. When a perceptual difference was introduced between the target and masker phrases by making one voiced and one whispered (diamonds: mixed conditions), performance was better or equal to that obtained when both phrases were voiced. This suggests that listeners are able to use differences in voicing to segregate the target talker from a speech masker.

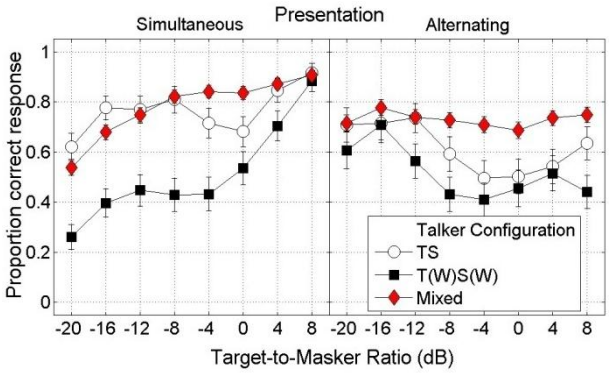

Figure 4: *Proportion correct color-number responses for a voiced target phrase presented simultaneously (left panel) or alternating (right panel) with a voiced speech masker (TS: circles), a whispered target presented along with a whispered speech masker (T(W)S(W):squares), and a voiced or whispered target talker presented with a whispered or voiced target (Mixed: diamonds).*

3.3. Effect of target talker in mixed configurations

Figure 5 shows the performance as a function of TMR in the mixed configuration: i.e., in conditions where a voiced target was presented with a simultaneous (left panel) or alternating (right panel) whispered speech masker (circles) or a whispered target was presented simultaneously or alternating with a voiced masker. From the results, it is clear that a whispered target – voiced masker is more intelligible at negative TMRs than a voiced target – whispered masker. One possible reason for the asymmetry is because of the fact that consonant-vowel energy ratio is higher in whispered speech which might favor the whispered speech when TMR in calculated relative to the voiced masker. These patterns of results are similar to those reported by [9] who reported that recognition scores based on SNR for syllables, consonants, and vowels were better when a whispered target was detected against a voiced masker than vice-versa.

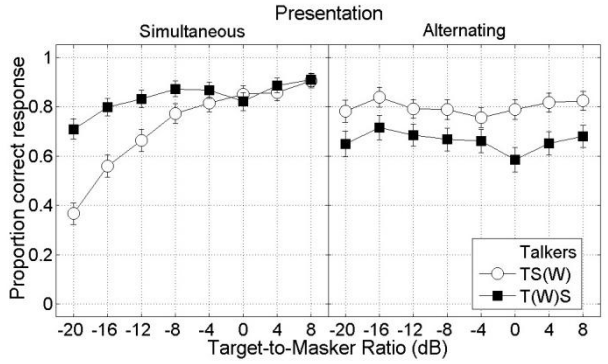

Figure 5: *Proportion correct color-number responses depicted for target and maskers that are presented simultaneously (left panel) or alternating (right panel). The circles depict performance with a voiced target and whispered masker. Squares depict intelligibility with a whispered target and voiced masker.*

The pattern of results is exactly the opposite in the sequential condition. In the sequential listening condition, listeners were better at identifying color and number of a

target phrase when the target was voiced and the masker was whispered. Results obtained in this experiment imply that pitch cues might be extremely important in perceptual segregation task, where listeners have to track a target signal that is interspersed with a noise-like masker.

3.4. Comparing TD and mixed talker configuration

Results from the current experiment suggest that listeners are best at segregating a mixture that comprises two voices varying in their voicing characteristics. The differentiation of talkers based on voicing could be an effective method to separate multiple talkers in a speech display. One question that still remains is whether or not the performance obtained with a whispered target-voiced masker is better compared to two voices that differ from each other based on sex (i.e., the TD listening configuration). Thus, a follow-up experiment was conducted to compare the performance of listeners in the following listening conditions; in the simultaneous listening condition, performance in the TD condition was compared to that obtained with a whispered target-voiced masker configuration. Since our data showed that performance in the sequential condition was best when a voiced target was presented with a whispered masker, that configuration was compared with a case where two different sex talkers were alternated. Those results are depicted in Figure 6. The circles depict performance in the TD configuration, whereas the squares depict performance in the mixed condition. It is clear that performance in the mixed condition is comparable to the TD condition at positive TMRs, but the use of two different sex talkers is a more effective strategy to improve intelligibility in a speech display comprising of two simultaneous talkers, especially at negative signal-to-noise ratios. For all TMRs in the alternating case, performance with a TD mixture is comparable if not better than the condition with a voiced target-whispered masker.

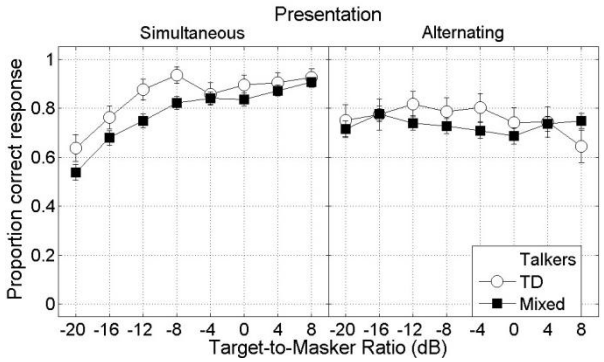

Figure 6: *Proportion correct color & number responses for a voiced target phrase presented simultaneously (left panel) or alternating (right panel) with a voiced speech masker (TD: circles). The squares in the simultaneous case represent performance in the T(W)S configuration, whereas the squares in the alternating case represents performance in the TS(W configuration.*

4. Conclusions

The current study investigated the importance of voicing in the identification of simultaneously or sequentially presented speech stimuli. Listeners can effectively segregate a whispered talker from a voiced talker when presented simultaneously, but performance degrades when the voiced masker is alternated with the whispered target. This suggests that fundamental frequency cues missing in whispered speech might be important in linking together elements of a target signal across time. The use of whispered speech could be a viable strategy in a speech displays, which require listeners to segregate two talkers, but is not as effective as using a different sex talker. Due to the performance benefit obtained for whispered speech over voiced speech in the presence of a noise masker, it could potentially afford an advantage in high noise environments.

5. Acknowledgements

This research was funded by a grant from the Air Force Office of Scientific Research (AFOSR).

6. References

[1] Traunmüller, H., and Eriksson, A. 2000. “Acoustic effects of variation in vocal effort by men, women, and children,” J. Acoust. Soc. Am. 107, 3438–3451.

[2] Schwartz, M. F. 1970. “Power spectral density measurements of oral and whispered speech,” J. Speech Hear. Res. 13, 445–446.

[3] Tartter, V. C. 1991. “Identifiability of vowels and speakers from whispered syllables,” Percept. Psychophys. 49, 365–372.

[4] Tartter, V. C. 1989. “What’s in a whisper?” J. Acoust. Soc. Am. 86, 1678–1683.

[5] Schwartz, M. F., and Rine, H. E. 1968. “Identification of speaker sex from isolated, whispered vowels,” J. Acoust. Soc. Am. 44, 1736–1737.

[6] Tartter, V. C., and Braun, D. 1994. “Hearing smiles and frowns in normal and whisper registers,” J. Acoust. Soc. Am. 96, 2101–2107.

[7] Scheffers, M.T.M., 1983. “Sifting vowels: auditory pitch analysis and sound segregation,” Unpublished doctoral thesis, University of Groningen.

[8] Lea, A., 1992. “Auditory modeling of vowel perception,” Unpublished doctoral thesis, University of Nottingham.

[9] Vestergaard, M. D. and Patterson, R. D. (2009). “Effects of voicing in the recognition of concurrent syllables (L).” J Acoust Soc Am. 126, 2860–2863.

[10] Griffin, A.M., Freyman, R. L., Padre, J., Oxenham, A. J. (2011). “Release from masking in whispered nonsense sentences.” Poster presented at a meeting of American Auditory Society, Scottsdale, AZ.

[11] Iyer, N., Brungart, D.S., and Simpson, B.D. (2007). “Effects of periodic masker interruption on the intelligibility of interrupted speech,”J. Acoust. Soc. Am. 122, 1693–1701.

[12] Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (2000). “A speech corpus for multitalker communications research,” J. Acoust. Soc. Am. 107, 1065–1066.

[13] Brungart, D. (2001b). “Informational and energetic masking effects in the perception of two simultaneous talkers,” J. Acoust. Soc. Am. 109, 1101–1109.